ADAM pour le Deep Learning

ACHIQ Aya, CLETZ Laura, EL MAZZOUJI Wahel

Octobre 2025

UNIVERSITÉ DE MONTPELLIER

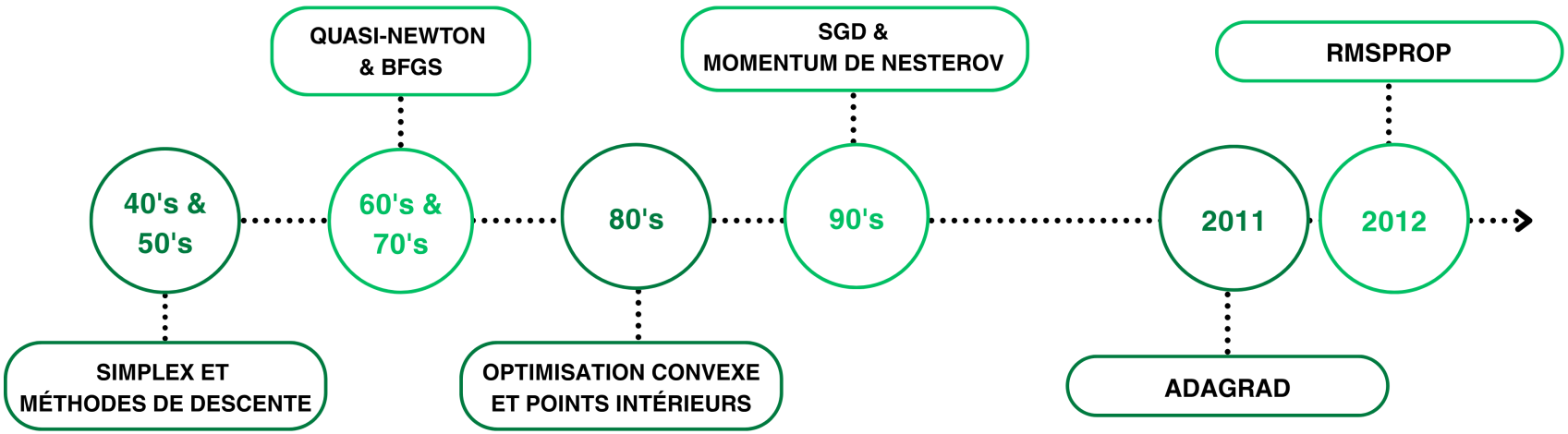
STATISTIQUE
SCIENCE DES DONNÉES
UNIVERSITÉ DE MONTPELLIER

FACULTÉ DES SCIENCES
DE MONTPELLIER

Sommaire

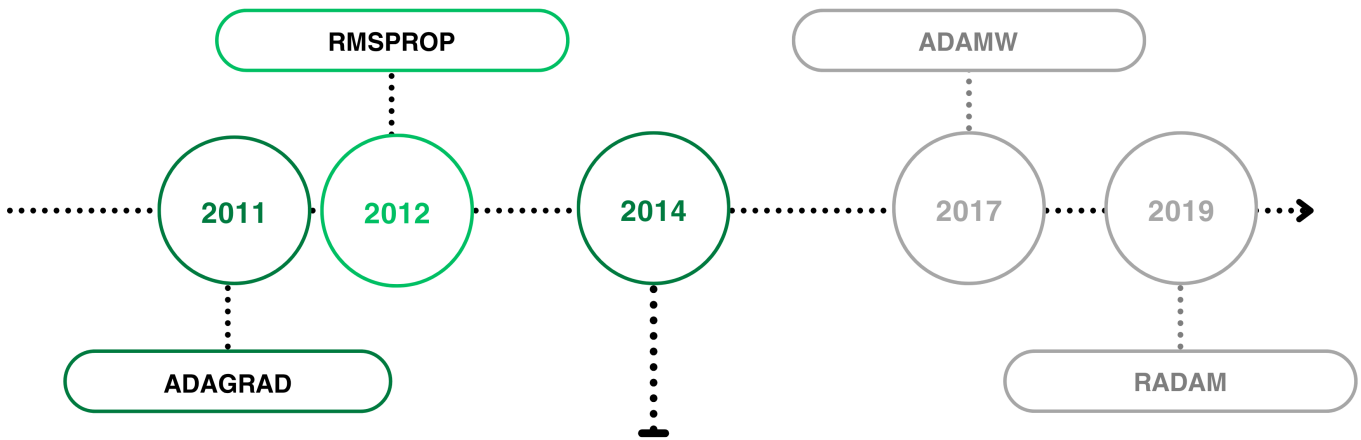
Optimisation en Deep Learning

- But : minimiser le fonction de **perte** $\mathcal{L}(\theta)$ pour un poids θ.
- Enjeux :
  - Convergence rapide ;
  - Stabilité numérique ;
  - Bonne généralisation.



Optimisation en Deep Learning

- But : minimiser le fonction de **perte** $\mathcal{L}(\theta)$ pour un poids θ.
- Enjeux :
  - Convergence rapide ;
  - Stabilité numérique ;
  - Bonne généralisation.

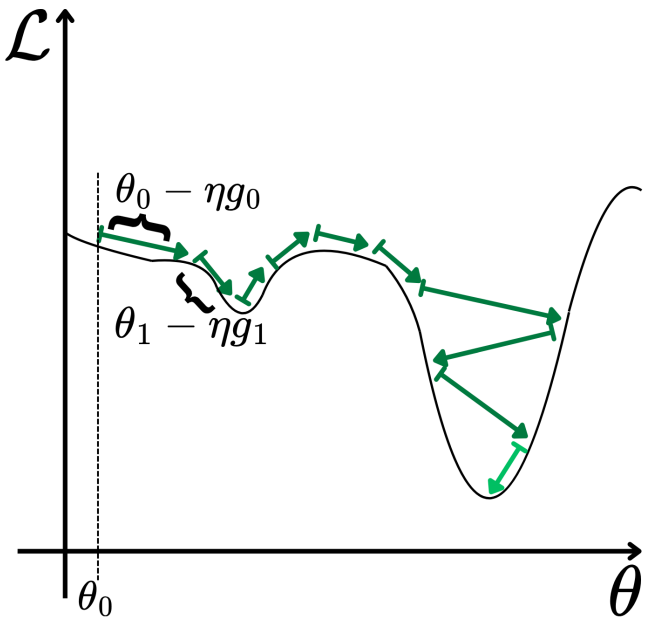


ADAM, Kingma et Ba, 2014

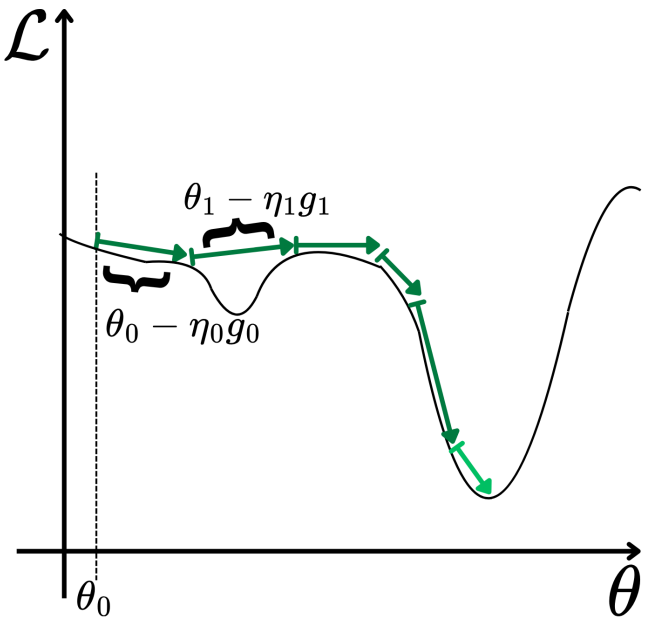
- Est-ce l'algorithme d'optimisation idéal pour le Deep Learning ?

Notions d'optimisation

- Notations :
  - Le gradient $g_t = \nabla_\theta \mathcal{L}(\theta_{t-1})$;
  - Le **learning rate** η : contrôle la taille des **pas** de mise à jour des paramètres, diffère suivant la méthode (Ruder, 2016) ;
  - Le **momentum** m_t : lissage de la trajectoire des gradients, calculé à partir de g_{t-1}.



Stochastic Gradient Descent



Root Mean Square Propagation

Principe de l'algorithme Adam

Idée clé

Adam = Adaptive Moment Estimation : combine le momentum et une adaptation du pas pour chaque paramètre.

- Moyennes mobiles :

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \quad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$$

- Mise à jour :

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon}$$

- Valeurs typiques : α=0.001, β_1=0.9, β_2=0.999.

Propriétés : invariance d'échelle, stabilité, faible sensibilité aux hyperparamètres.

Objectifs et limites d'Adam

Objectifs :

- Améliorer la convergence en combinant momentum et taux d'apprentissage adaptatif ;
- Stabiliser l'apprentissage, même lorsque les gradients sont bruités ;
- Réduire la sensibilité aux hyperparamètres grâce à l'adaptation automatique des pas.

Limites :

- Moindre généralisation que SGD Wilson et al., 2017 ;
- Convergence vers $\|w\|_\infty$ minimale ;
- Compromis entre vitesse et généralisation.

Expérimentation : Méthodologie

Credit Card

- Taille : **10k / 100k**
- Features : **2**
- Classes : Déséquilibrées
- Réseau : [2→8→4→1]

→ Dataset large, simple

Heart Disease

- Taille : **1025**
- Features : **13**
- Classes : Équilibrées
- Réseau : [13→16→8→1]

→ Dataset petit, complexe

Optimiseurs : Adam, SGD, Adagrad, RMSprop
Learning rates : 0.001 (Adam/RMSprop), 0.01 (SGD/Adagrad)

Credit Card Fraud Detection

n = 10 000

Optimiseur	Loss	Erreur (%)
Adam	**0.0057**	0.23
SGD	0.0224	0.23
Adagrad	0.0271	0.23
RMSprop	0.0167	0.23

n = 100 000

Optimiseur	Loss	Erreur (%)
Adam	**0.0033**	**0.09**
SGD	0.0046	0.14
Adagrad	0.0042	0.10
RMSprop	0.0045	0.10

Observations

- **n=10k :** Tous à 0.23% d'erreur, mais la *loss* révèle qu'Adam optimise mieux (0.0057)
- **n=100k :** Adam se détache nettement (0.09% d'erreur, loss 0.0033)

Heart Disease

n = 1 025

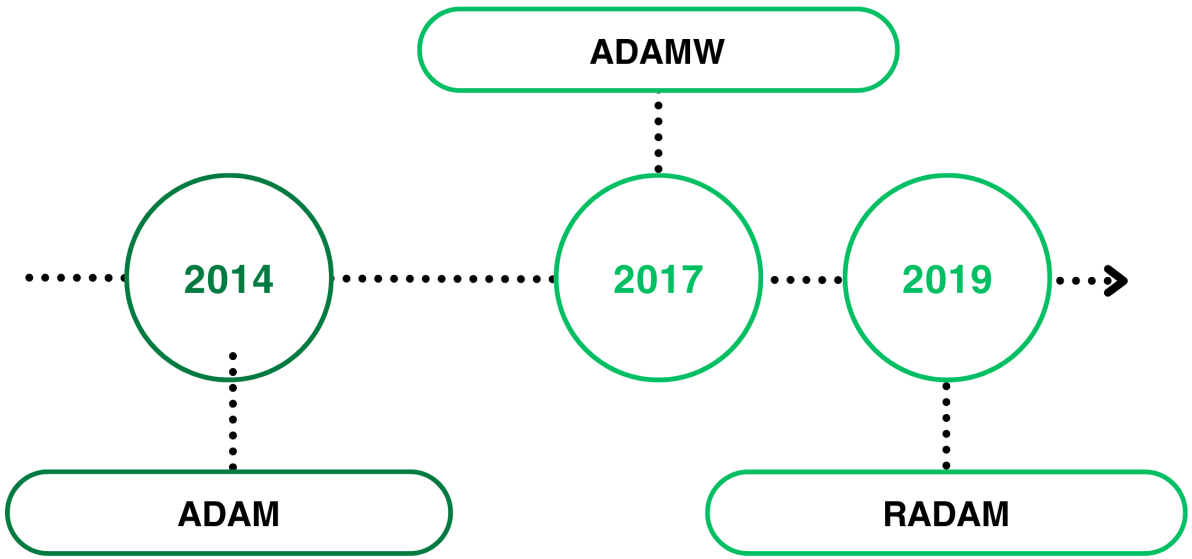
Optimiseur	Test Loss	Test Accuracy	Erreur (%)
Adam	0.2196	0.9318	6.82
SGD	0.3425	0.8084	**19.16**
Adagrad	0.2946	0.8864	11.36
RMSprop	**0.2250**	**0.9416**	**5.84**

Observations

- **Gros écarts** : RMSprop meilleur (5.84%), Adam proche (6.82%), SGD s'effondre (19.16%)
- Dataset petit (1k) + haute dimension (13D) → favorise les méthodes adaptatives

Conclusion

- Avantages d'Adam :
  - rapidité à l'initialisation ;
  - convergence rapide par momentum ;
  - stabilité numérique en présence de gradients bruités ;
  - adaptation automatique des pas.
- Limites d'Adam :
  - faible capacité de généralisation (surapprentissage) ;
  - dépendance de la qualité des données.



Références

Kingma, Diederik P. et Jimmy Lei Ba (2014). « **Adam : A Method for Stochastic Optimization** ». In : url : https://arxiv.org/abs/1412.6980.

Ruder, Sebastian (2016). « **An overview of gradient descent optimization algorithms** ». In : url : https://arxiv.org/pdf/1609.04747.

Wilson, Ashia C. et al. (2017). « **The Marginal Value of Adaptive Gradient Methods in Machine Learning** ». In : url : https://arxiv.org/abs/1705.08292.